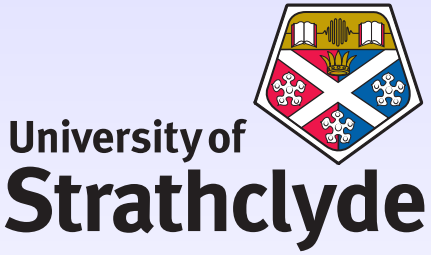
Analysis of Some Simple Stochastic Models and Algorithms in Biology and Chemistry

Des Higham
Department of Mathematics and Statistics
University of Strathclyde
djh@maths.strath.ac.uk



Motivation

Chemical kinetics is being used to model processes inside the cell.

Different modeling regimes can be used–**how do they compare**?

We will look at:

- Mean exit times for jump versus diffusion.
- Moment accuracy of hybrid discrete/continuous models.
- Relative noise strengths in hierarchies of gene regulation models.

Markov Jump Versus Diffusion

In many applications,including

- **chemistry**,
- **cell biology**,
- **population dynamics**,
- **epidemiology**,

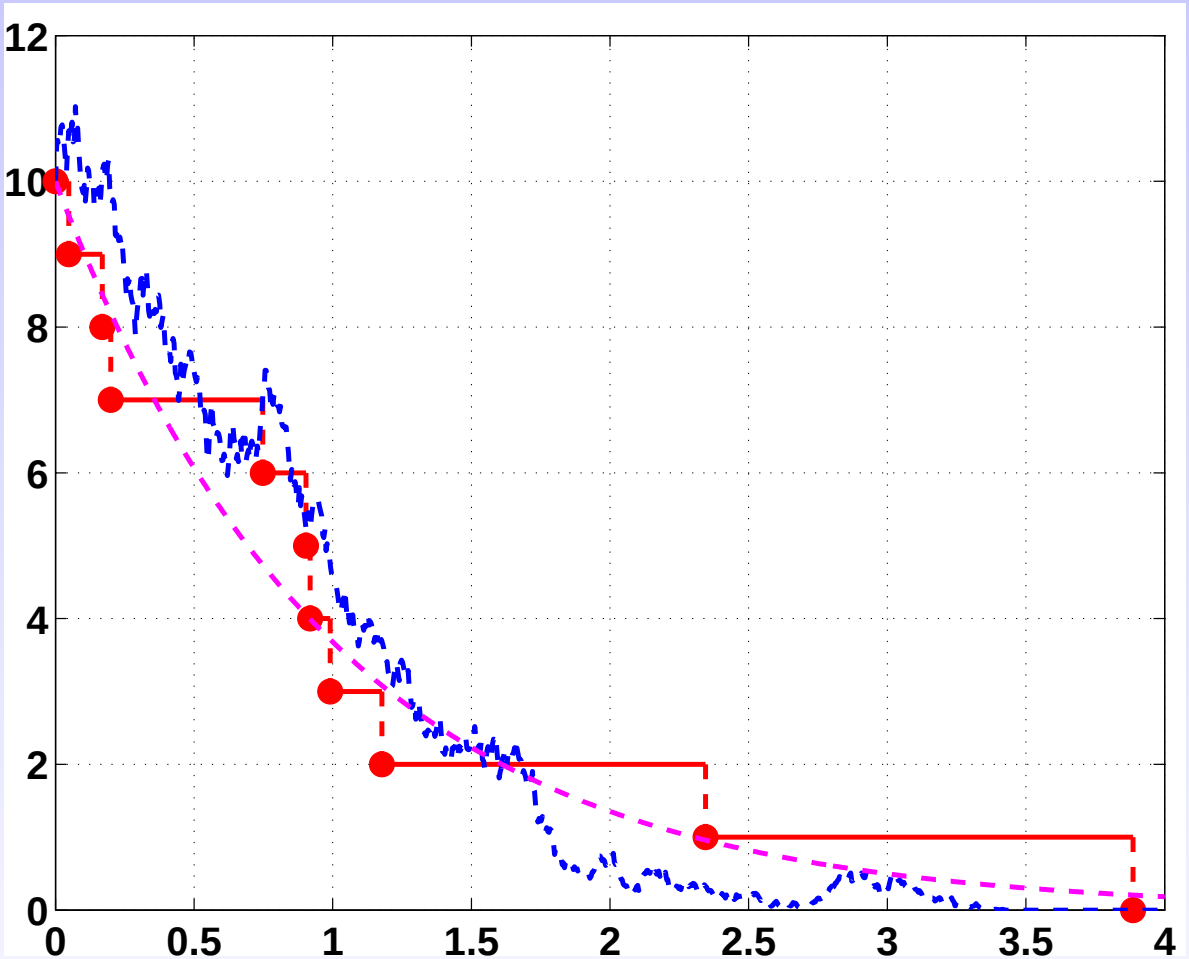
we can model at different levels:

E.g.

- **CME** (Jump): what is probability that we have 237 proteins at time t?
- **CLE** (Diffusion): what is probability that protein concentration is between 2.7 and 3.1 at time t?
- **RRE** (mass action ODE): what is protein concentration at time t?

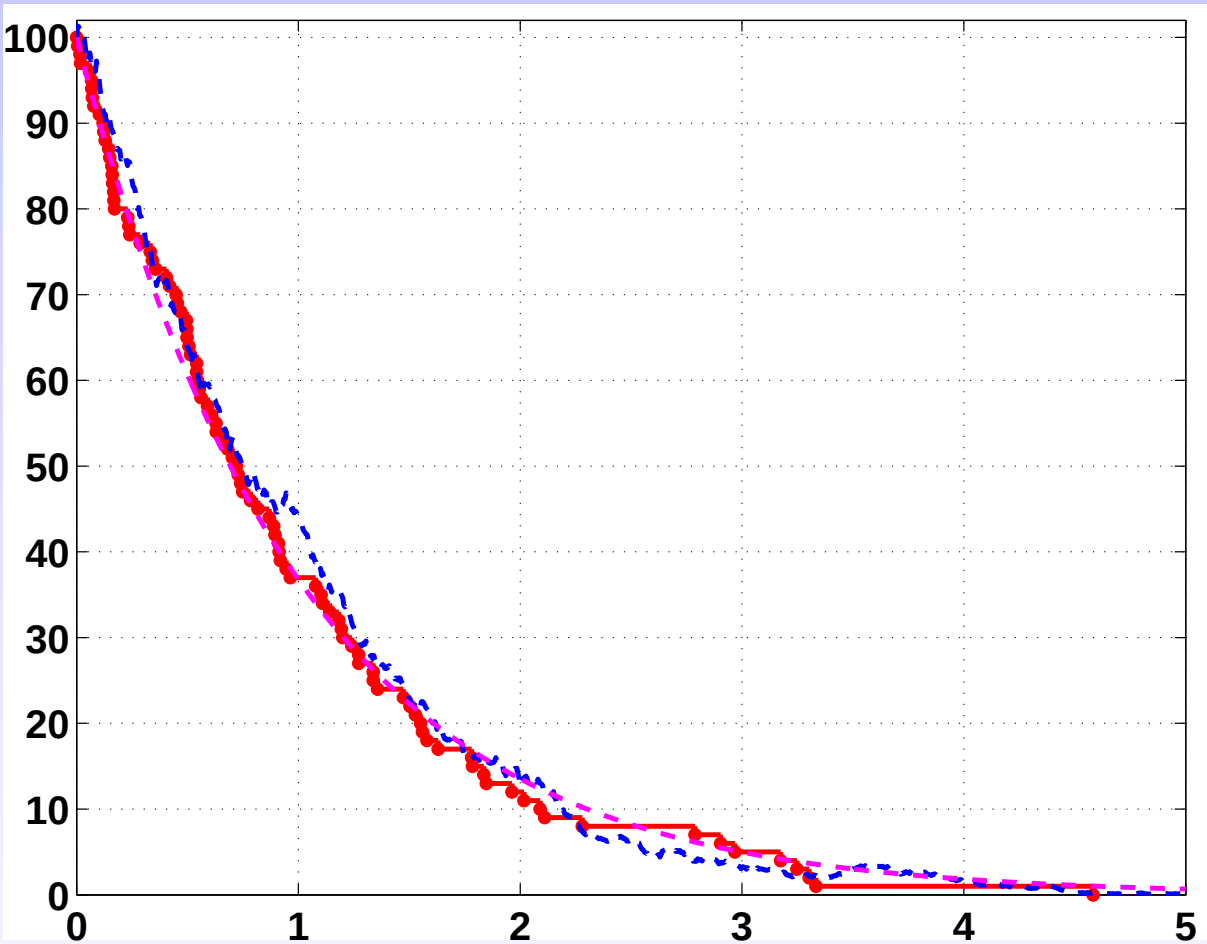
These modeling regimes 'converge' when the population size increases . . . . . how do we quantify this?

Example: $S \xrightarrow{c=1} \emptyset$, start with 10 molecules



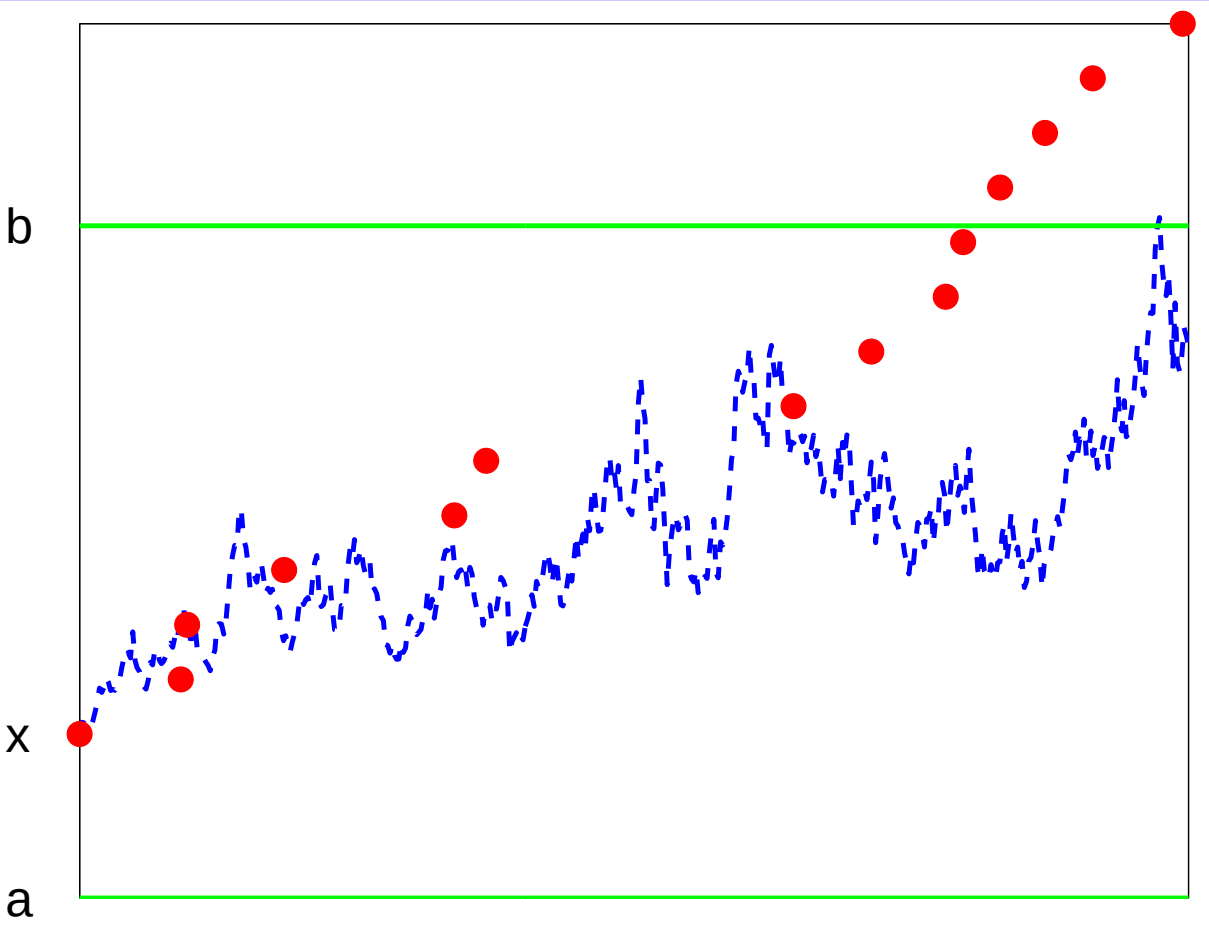
CME CLE RRE

Example: $S \xrightarrow{c=1} \emptyset$, start with 100 molecules



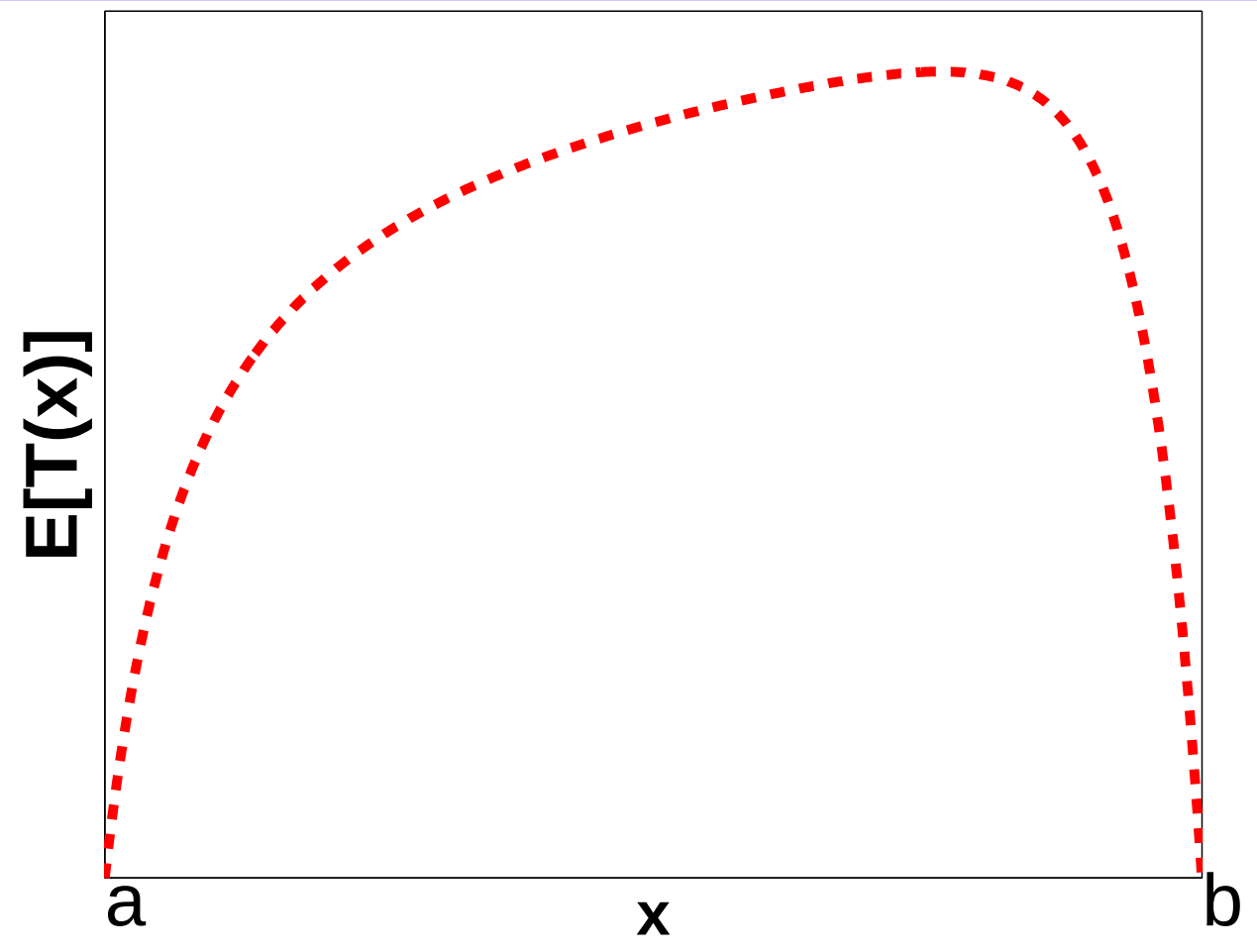
CME CLE RRE

Focus first on **mean hitting time**



$$T(x) := \inf(t : \mathbf{Z}(t) = a \quad \text{or} \quad \mathbf{Z}(t) = b, \quad \text{given} \quad \mathbf{Z}(0) = x)$$

Focus first on **mean hitting time**



Markov jump/birth & death process, $\mathbf{Z}(t)$

Discrete states $\{0, 1, 2, \ldots, M\}$, with 0 and M absorbing:

$$\begin{aligned} P\left(\mathbf{Z}(t+h) = i+1 \mid \mathbf{Z}(t) = i\right) &= B_i h + o(h), \\ P\left(\mathbf{Z}(t+h) = i-1 \mid \mathbf{Z}(t) = i\right) &= D_i h + o(h), \\ P\left(\mathbf{Z}(t+h) = i \mid \mathbf{Z}(t) = i\right) &= 1 - (B_i + D_i)h + o(h). \end{aligned}$$

Here, $B_0 = D_0 = B_M = D_M = 0$.

Starting at state $\mathbf{Z}(0) = j$, the expected time to be absorbed into state 0 or M is given by U_j, where

$$\begin{bmatrix} (B_1+D_1) & -B_1 & & & & \\ -D_2 & (B_2+D_2) & -B_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -B_{M-2} \\ & & & & -D_{M-1} & (B_{M-1}+D_{M-1}) \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ \vdots \\ U_{M-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

Numerical Analysis Viewpoint

Linear system can be written, for $1 \leq i \leq M-1$,

$$\frac{B_i + D_i}{2}\left(U_{i+1} - 2U_i + U_{i-1}\right) + \left(B_i - D_i\right)\frac{U_{i+1} - U_{i-1}}{2} = -1$$

Standard **finite differences** on the 2 point BVP ODE

$$\frac{B(x) + D(x)}{2}u''(x) + \left(B(x) - D(x)\right)u'(x) = -1,$$

with $u(a) = u(b) = 0$

Here $b - a = M$ and we have $\Delta x = 1$

Interesting regime is $M \to \infty$

Diffusion Approximation

SDE:

$$\begin{aligned} d\mathbf{y}(t) &= (B(\mathbf{y}(t)) - D(\mathbf{y}(t)))\, dt \\ &\quad + \sqrt{B(\mathbf{y}(t))}\, d\mathbf{W}_1(t) - \sqrt{D(\mathbf{y}(t))}\, d\mathbf{W}_2(t) \end{aligned}$$

Let $w(x) := \mathbb{E}[T(x)]$ be the expected first time to hit a or b, given that $\mathbf{y}(0) = x$.

Then $w(x)$ satisfies the same 2 point BVP ODE.

Want to show that this ODE 'converges' to the finite difference scheme.

Focus on specific examples ...

Production from a source

$$\emptyset \xrightarrow{k} X \qquad B_i = k \quad \text{and} \quad D_i = 0$$

Mean hitting times:

Jump process

$$\frac{b-x}{k}$$

Diffusion process

$$\frac{1}{k}\left[\frac{-e^{-2x}+e^{-2a}}{-e^{-2b}+e^{-2a}}\left(\frac{-e^{-2b}}{2}\left(e^{2b}-e^{2x}\right)+b-x\right)\right.$$
$$\left.+\left(1-\frac{-e^{-2x}+e^{-2a}}{-e^{-2b}+e^{-2a}}\right)\left(a-x+\frac{e^{-2a}}{2}\left(e^{2x}-e^{2a}\right)\right)\right]$$

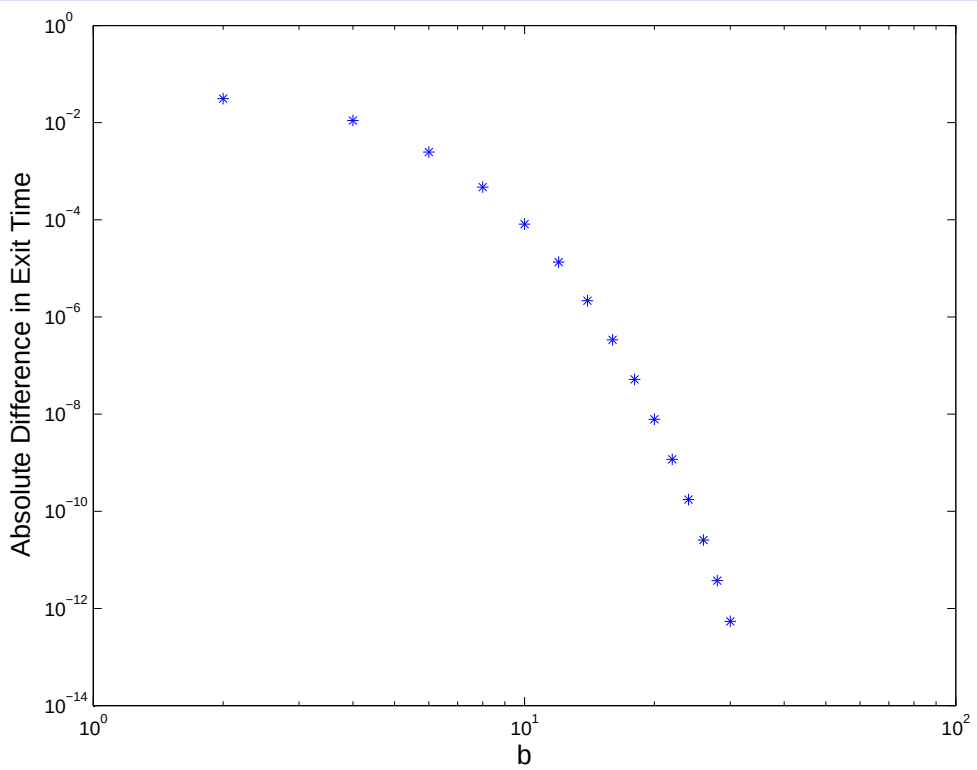
Convergence: fix $a = 0$ and let $b \to \infty$

With $x = \alpha b$ for fixed $\alpha \in (0, 1)$, we have

$$|\text{Jump} - \text{Diffusion}| \leq C\, e^{-b \min\{2(1-\alpha), \alpha\}}$$

where C is independent of b.

Example: $k = 5$, $a = 0$, $\alpha = \frac{1}{2}$:



Production

$$\emptyset \xrightarrow{c\,X} X \qquad B_i = i \quad \text{and} \quad D_i = 0$$

Mean hitting times:

Jump process

$$\frac{1}{c}\sum_{s=x}^{b-1}\frac{1}{s}$$

Diffusion process

$$\frac{1}{c}\left(\frac{e^{-2x}-e^{-2a}}{e^{-2b}-e^{-2a}}\left(-e^{-2b}\int_x^b \frac{e^{2l}}{l}dl + \ln b - \ln x\right)\right.$$
$$\left. + \left[1 - \frac{e^{-2x}-e^{-2a}}{e^{-2b}-e^{-2a}}\right]\left(\ln a - \ln x + e^{-2a}\int_a^x \frac{e^{2l}}{l}dl\right)\right)$$

Convergence

With $x = \alpha b$ for fixed $\alpha \in (0, 1)$, we have

$$|\text{Jump } (\text{with } a = 0) - \lim_{a \searrow 0} \text{Diffusion}| \leq Cb^{-2}$$

where C is independent of b.

Proof Uses the expansions

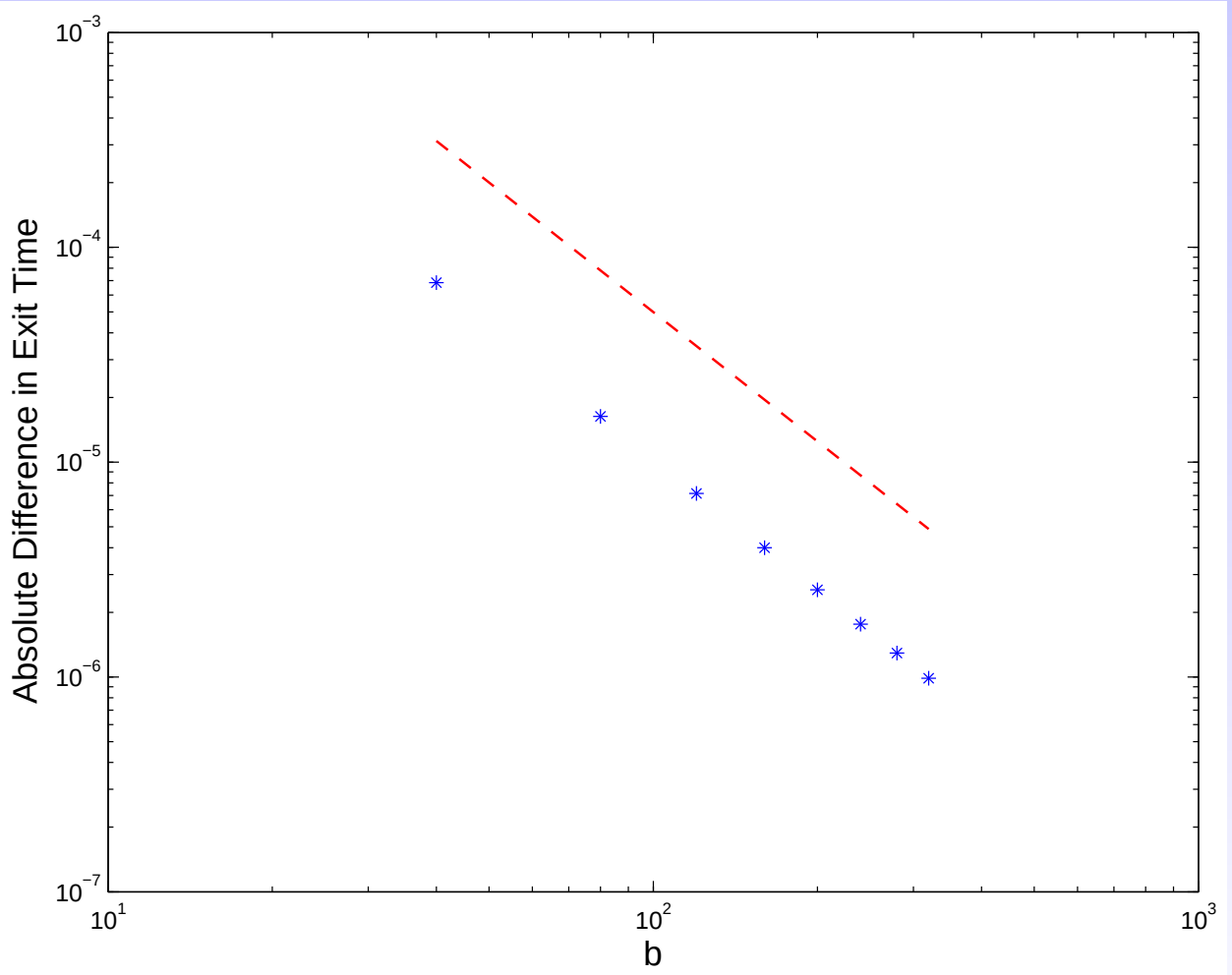
$$\sum_{s=1}^{n} \frac{1}{s} = \ln n + \gamma + \frac{1}{2n} + O(n^{-2}), \qquad \text{as } n \to \infty,$$

$$\int_{-\infty}^{x} \frac{e^t}{t}\, dt = \ln x + \gamma + o(1), \qquad \text{as } x \to 0,$$

where the **Euler-Mascheroni constant** $\gamma = 0.5772\ldots$, and

$$\int_{-\infty}^{x} \frac{e^t}{t}\, dt = \frac{e^x}{x}\left(1 + \frac{1}{x} + O(x^{-2})\right), \qquad \text{as } x \to \infty.$$

Example, $c = 5$, $a = 10^{-3}$, $\alpha = \frac{1}{2}$



Degradation

$$X \xrightarrow{c} \emptyset \qquad B_i = 0 \quad \text{and} \quad D_i = i$$

Mean hitting times:

Jump process

$$\frac{1}{c}\sum_{s=a+1}^{x}\frac{1}{s}$$

Diffusion process

$$\frac{e^{2x}-e^{2a}}{e^{2b}-e^{2a}}\left(\frac{1}{c}\left(e^{2b}\int_x^b \frac{e^{-2l}}{l}dl - \ln b + \ln x\right)\right)$$
$$+\left[1-\frac{e^{2x}-e^{2a}}{e^{2b}-e^{2a}}\right]\left(\frac{1}{c}\left(\ln x - \ln a - e^{2a}\int_a^x \frac{e^{-2l}}{l}dl\right)\right)$$

Convergence

With $x = \alpha b$ for fixed $\alpha \in (0, 1)$, we have

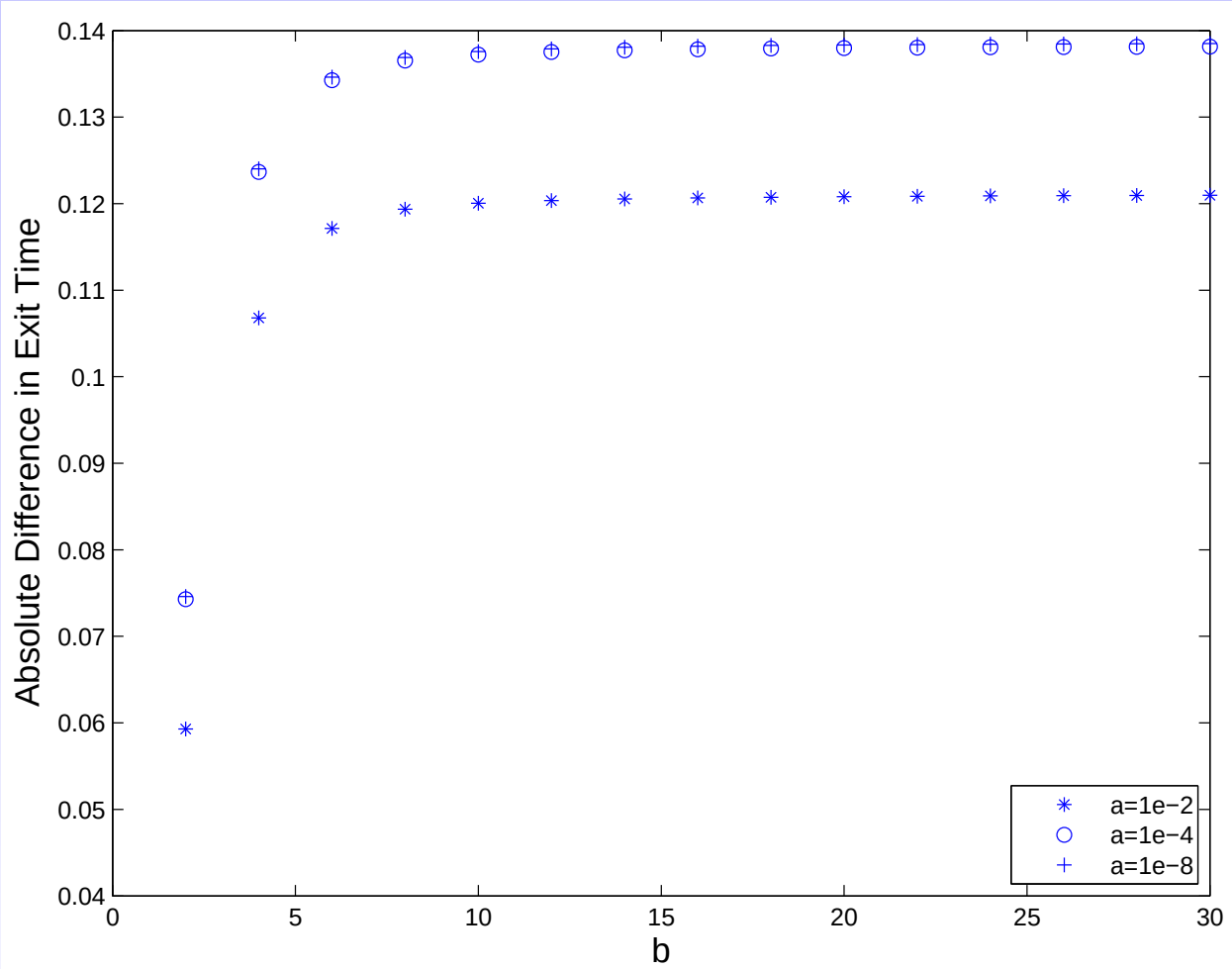
$$\lim_{a \searrow 0} \lim_{b \to \infty} (\text{Jump} - \text{Diffusion}) = \frac{-\ln 2}{c}$$

Proof Uses asymptotic expansions for

$$E_1(x) = \int_x^\infty \frac{e^{-t}}{t} dt, \qquad x > 0.$$

Note: the actual hitting times grow like $\ln(b)$, so we have **relative convergence** like $O(1/\ln b)$.

Example, $c = 5$, $\alpha = \frac{1}{2}$, $a = 10^{-2}, 10^{-4}, 10^{-8}$



$$\frac{\ln 2}{5} = 0.1386\ldots$$

Discussion

This approach of **expanding exact solutions** breaks down for more complicated scenarios. E.g.

$$\emptyset \xrightarrow{k} X \xrightarrow{cX} \emptyset$$

involves integrals of the incomplete Gamma function.

Is there a **general framework** for analysing finite difference schemes in this non-standard context?
At best, convergence is relative not absolute.

Technical Issue

Looking at **mean hitting times** is practically relevant, and avoids pitfalls that can arise through the SDE breaking down.
E.g. consider the **reversible isometry**

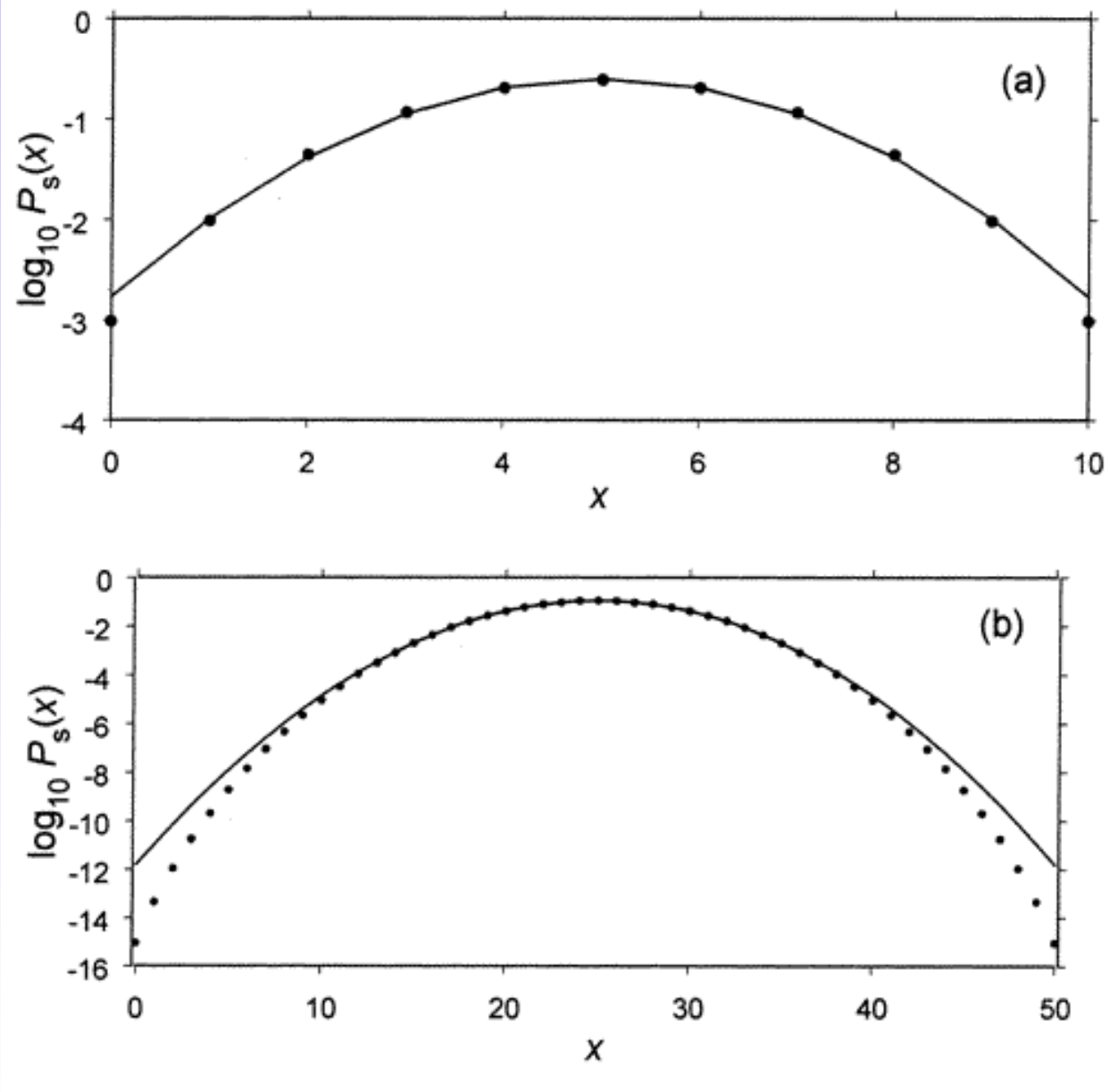
$$X_1 \underset{c_2 X_2}{\overset{c_1 X_1}{\rightleftharpoons}} X_2$$

Gillespie, J. Phys. Chem. **2002**
The Chemical Langevin and Fokker–Planck Equations for the Reversible Isomerization Reaction

SDE:

$$d\mathbf{Y} = (-c_1\mathbf{Y} + c_2(K - \mathbf{Y}))\,dt - \sqrt{c_1\mathbf{Y}}d\mathbf{W}_1 + \sqrt{c_2(K-\mathbf{Y})}d\mathbf{W}_2$$

$c_1 = c_2$ steady state distributions?



Chemical Kinetics (Gillespie 1976)

N chemical species, M types of reaction (e.g. $A + B \to C$)
State vector

$$\mathbf{X}(t) = \begin{bmatrix} X_1(t) \\ X_2(t) \\ \vdots \\ X_N(t) \end{bmatrix}, \qquad \mathbf{X}(0) = \mathbf{X}_0$$

Each reaction, $1 \leq j \leq M$, is described by

- a stoichiometric vector $\nu_j \in \mathbb{R}^N$ such that $\mathbf{X}(t) \mapsto \mathbf{X}(t) + \nu_j$,
- a propensity function $a_j(\mathbf{X}(t))$ such that the prob. of this reaction taking place over time $[t, t + dt)$ is $a_j(\mathbf{X}(t))\, dt$

CME

Discrete state space, continuous time Markov chain.

Let $P(\mathbf{x}, t)$ be the prob. that $\mathbf{X}(t) = \mathbf{x}$.

$$\frac{dP(\mathbf{x}, t)}{dt} = \sum_{j=1}^{M} (a_j(\mathbf{x} - \nu_j)P(\mathbf{x} - \nu_j, t) - a_j(\mathbf{x})P(\mathbf{x}, t))$$

Gillespie's stochastic simulation algorithm gives a way to compute realisat ions of $(t, \mathbf{X}(t))$.

Takes account of every reaction $\Rightarrow$ **expensive**.

CLE

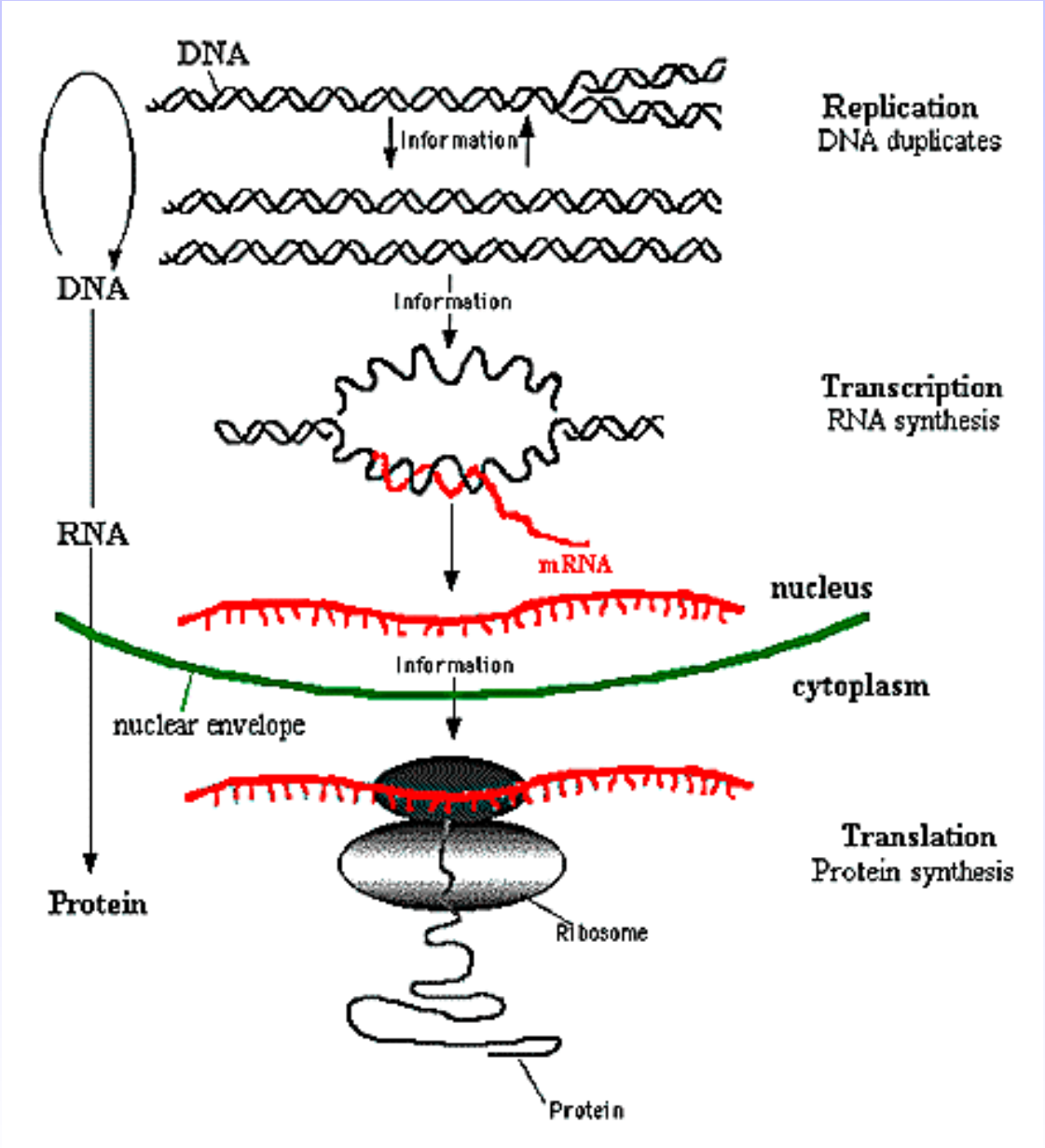
SDE in $\mathbb{R}^N$.

$$d\mathbf{Y}(t) = \sum_{j=1}^{M} \nu_j a_j(\mathbf{Y}(t))dt + \sum_{j=1}^{M} \nu_j \sqrt{a_j(\mathbf{Y}(t))}\, d\mathbf{W}_j(t)$$

Euler–Maruyama computes approximate realisations of $(t, \mathbf{Y}(t))$.

(Switching off the noise gives the RRE.)

Central Dogma of Cell Biology



Gene Transcription Model

Raser & **O'Shea**, **Science**, 2004:

$$\left.\begin{array}{l} D_i \xrightarrow{k_a} D_i^\star \\ D_i \xleftarrow{k_b} D_i^\star \\ D_i^\star \xrightarrow{k_r} D_i^\star + M \end{array}\right\} \quad 1 \le i \le m$$

and

$$\begin{array}{rcl} M & \xrightarrow{k_p} & M + P \\ M & \xrightarrow{\gamma_r} & \emptyset \\ P & \xrightarrow{\gamma_p} & \emptyset \end{array}$$

Hybrid approach (e.g. **Paszek**, **Bull. Math. Biol.**, 2007): treat the D_i and $D_i^\star$ as discrete and M and P as continuous: **switching ODE**

Alternative Hybrid Model

Let $\mathbf{r}(t)$ denote the number of active genes at time t.
Then $\mathbf{r}(t)$ takes values in $\{0, 1, 2, 3, \ldots, m\}$ driven by a continuous time Markov chain.
Using the CLE framework for the remaining reactions we get a **switching SDE**:

$$d\begin{bmatrix} \mathbf{M} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} k_r \mathbf{r} & -\gamma_r \mathbf{M} \\ k_p \mathbf{M} & -\gamma_p \mathbf{P} \end{bmatrix} dt$$

$$+ \begin{bmatrix} \sqrt{k_r \mathbf{r}} & -\sqrt{\gamma_r \mathbf{M}} & 0 & 0 \\ 0 & 0 & \sqrt{k_p \mathbf{M}} & -\sqrt{\gamma_p \mathbf{P}} \end{bmatrix} \begin{bmatrix} d\mathbf{W}_1 \\ d\mathbf{W}_2 \\ d\mathbf{W}_3 \\ d\mathbf{W}_4 \end{bmatrix}$$

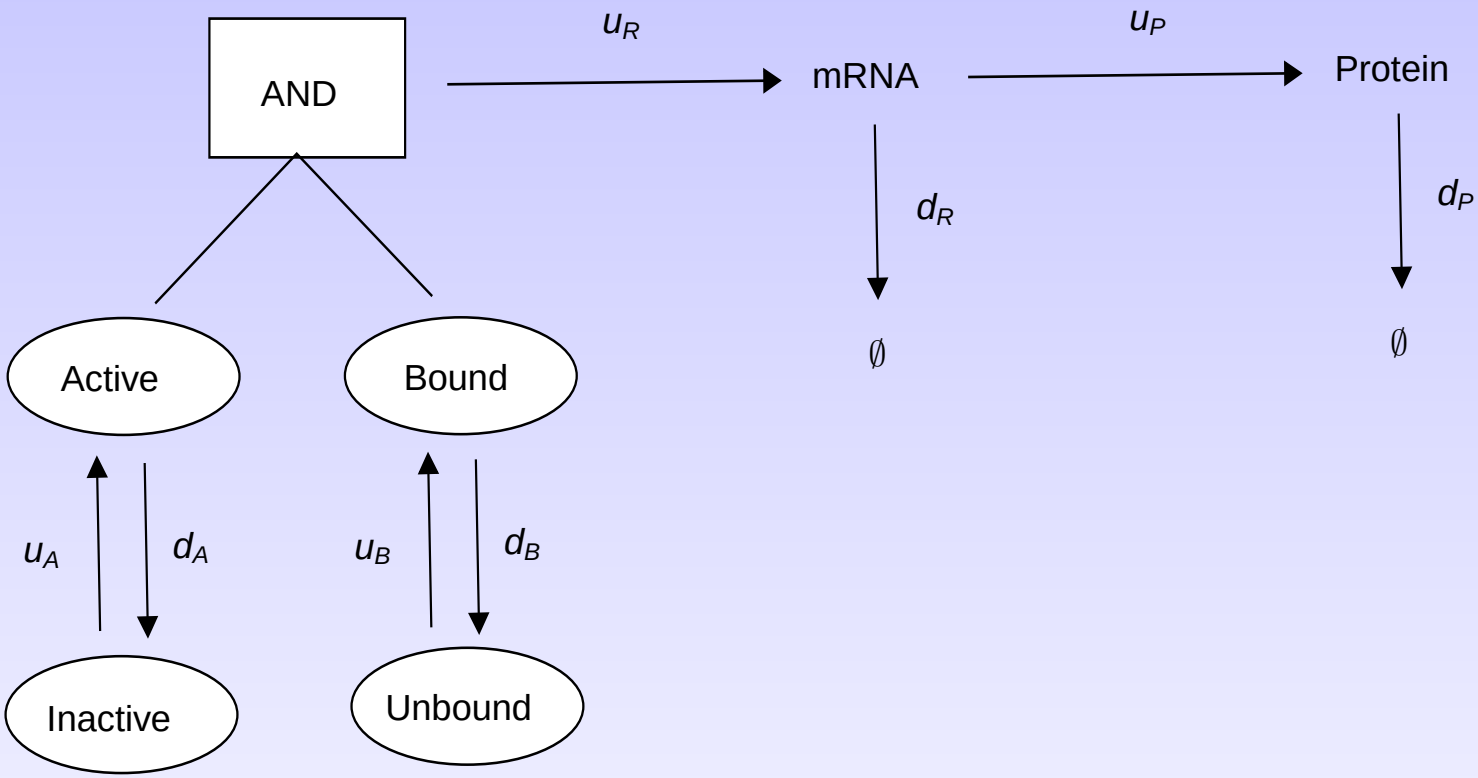
Means, Variances and Correlations

There is a generalized version of Ito's Lemma for switching SDEs (**Mao** and **Yuang**, 2006).

Using this:

New Result $\mathbb{E}[\mathbf{r}]$, $\mathbb{E}[\mathbf{M}]$, $\mathbb{E}[\mathbf{P}]$, $\mathbb{E}[\mathbf{Mr}]$, $\mathbb{E}[\mathbf{Pr}]$, $\mathbb{E}[\mathbf{MP}]$, $\mathbb{E}[\mathbf{r}^2]$, $\mathbb{E}[\mathbf{M}^2]$ and $\mathbb{E}[\mathbf{P}^2]$ for the hybrid model match those for the full CME.

“Two Switch” Model



We can write this as a first order network, and obtain ODEs for first and second moments.

“Two Switch” Model: Summary of Results

Variance can either increase or decrease when we add one or two switches, but

“Two Switch” Model: Summary of Results

Variance can either increase or decrease when we add one or two switches, but

at **steady state**:

“Two Switch” Model: Summary of Results

Variance can either increase or decrease when we add one or two switches, but

at **steady state**:

- **noise strength** for **one or two switches** is always **greater than for no switch**, yet

“Two Switch” Model: Summary of Results

Variance can either increase or decrease when we add one or two switches, but

at **steady state**:

- **noise strength** for **one or two switches** is always **greater than for no switch**, yet
- **two switches may be more or less noisy than one**.

“Two Switch” Model: Summary of Results

Variance can either increase or decrease when we add one or two switches, but

at **steady state**:

- **noise strength** for **one or two switches** is always **greater than for no switch**, yet
- **two switches may be more or less noisy than one**.

These steady state results **do not reflect the transient behaviour**.

“Two Switch” Model: Summary of Results

Variance can either increase or decrease when we add one or two switches, but

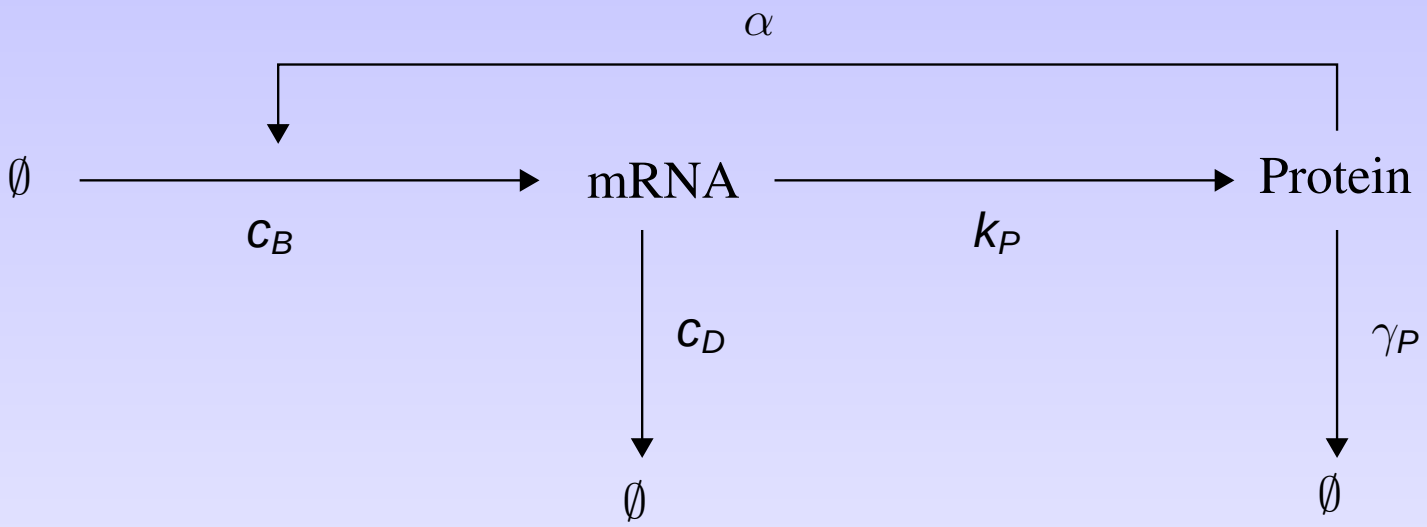
at **steady state**:

- **noise strength** for **one or two switches** is always **greater than for no switch**, yet
- **two switches may be more or less noisy than one**.

These steady state results **do not reflect the transient behaviour**.

Hybrid switch plus diffusion model correctly reproduces first and second moments.

Autoregulation



Consider the case where the protein linearly enhances its own production rate.

Autoregulation: Summary of Results

First and second moments (and correlations) of mRNA and protein **increase monotonically** with feedback strength, α.

Autoregulation: Summary of Results

First and second moments (and correlations) of mRNA and protein **increase monotonically** with feedback strength, α.

Not true for **variance** or **noise strength**.

Autoregulation: Summary of Results

First and second moments (and correlations) of mRNA and protein **increase monotonically** with feedback strength, α.

Not true for **variance** or **noise strength**.

At steady state (stable for $\alpha < (c_D\gamma_P)/(c_B\kappa_P)$) :

Autoregulation: Summary of Results

First and second moments (and correlations) of mRNA and protein **increase monotonically** with feedback strength, α.

Not true for **variance** or **noise strength**.

At steady state (stable for $\alpha < (c_D\gamma_P)/(c_B\kappa_P)$) :
variance and **noise strength** increase monotonically with α

Autoregulation: Summary of Results

First and second moments (and correlations) of mRNA and protein **increase monotonically** with feedback strength, α.

Not true for **variance** or **noise strength**.

At steady state (stable for $\alpha < (c_D\gamma_P)/(c_B\kappa_P)$) :
variance and **noise strength** increase monotonically with α

Hybrid switch plus diffusion model correctly reproduces first and second moments.

Summary

What's new?

Rigorous results on mean hitting times and first/second moments for simple models.

- CLE and CME **mean hitting times** don't match well
- **CLE** formulation can **break down**
- **ODE + switch** underestimates the variance
- **Diffusion + switch** gets it right
- **Extra switching** or **autoregulation** **increases the noise strength**

What's Next?

- Spatial effects (subdiffusion), delays, cell growth
- Other types of regulation
- Multiscale simulation algorithms
- Inference